STAT 315 Assignment 3

May 25, 2014

Lab: 12 May. Due Date: 16 May, 4pm. Submit via Learn. Maximum possible marks: 15.
Save the files `Xenon.txt` and `Nordic.txt` and the SAS file `a3code.sas` in the folder `P:\stat315`.

1. This question builds on your knowledge of SAS from Assignments 1 and 2. It is important to know how to interpret SAS output in a way that would make sense to a non-statistician.

   The file `Xenon.txt` contains measurements of pressure (kPa), temperature (millions of degrees celsius) and volume (cubic metres) for samples of 150g of Xenon from a star.

   (a) Run the SAS script to load in the data, and perform the regression of T on P and V. The R^2 is quite high. Do you have any reservations about the assumptions of linear regression here? Which plot or plots look(s) suspicious? (2)

   Yes, I have reservations [1 mark]. The plot of the residuals versus the predicted values looks like a U-shape. The spread of residuals varies with the predicted value of y. As we know, this is called heteroskedasticity. [1 mark]

   (b) Run the `proc gam` code and look at the output. The model fitted by SAS is

   $$T = \beta_0 + \beta_1 P + \beta_2 V + \beta_3 s(P) + \beta_4 s(V)$$

   where the $s(P)$ and $s(V)$ are smoothing terms. Only one of these terms is significantly different from zero at the $\alpha = 0.05$ level. Which one? (1)

   From the output:

   Spline(P) 3.00000 1.985866 5.2241 0.1561

   Spline(V) 3.00000 3.947855 10.3854 0.0156

   It is Spline(V) which has the significant p–value. That is $s(V)$. [1 mark]

   (c) Now create a new data set `Xenon2` with a new variable `PV`, defined as the product $P \times V$. (Look at the SAS code for Assignment 1 to see how to do this.) Run another regression with `PV` added to the linear model from part (a). Give the SAS code for this part in your solution. (2)

   ```
   data Xenon2;
   set Xenon;
   PV = P*V;
   run;

   proc reg data=Xenon2;
   model T = P V PV;
   run;
   ```

   (d) Using only the SAS output, which of the three models do you think is the best? How would you explain your choice to a non-statistician scientist? Which plot or plots would you show them? (3)

   The third model is the best [1 mark]. The plot of predicted value versus actual value is a perfect straight line. This model, with only three variables, fits the data almost perfectly. It looks like

the assumptions of linear regression are not violated here, and we got an R^2 close to 1. In non-statistician language, adding the extra term gives us a model which fits the data very well. We should be cautious about making claims about how our model might generalise to new data, but it looks very good.

2. The file `Nordic.txt` contains the result of the Sochi 2014 Nordic Combined 10k/Normal Hill event. The competition is decided by who performs the best in a combination of ski jumping and cross-country skiing. The variable `SkiJump` is the ski jump score and `CrossCountry` is the cross-country time in seconds. Source: `http://www.sochi2014.com/en/nordic-combined-ind-gund-nh-10-km-cross-c-free-race`

   (a) One way of combining the scores is to use the first principal component. Why might this be a good idea? (1)

   We want to say who is best by extracting a single number from two numbers. It is logical that we would want to do this in such a way that the competitors are as spread out as possible, which is the same as taking the first principal component. [1 mark]

   (b) If the competitors were ranked based on the first principal component, who would have won the bronze medal? (1)

   The person with the third highest value of Prin1 is the one in 6th place. Printing the data set reveals that this is Johannes RYDZEK of Germany [1 mark].

   (c) What do you think the *second* principal component represents? Are the data adequately summarized by one principal component? (2)

   The second principal component is 0.7*SJ + 0.7*CC. Since lower CC is better, this represents being good at Ski Jumping and bad at Cross-country. A high value of the 2nd Principal Component means that you are a better jumper than skier [1 mark].

   The data are not adequately summarised by one PC as only 50% of the variance is in the first PC. 1 mark].

   (d) The IOC wants to introduce a new snowmobile half-pipe event and is considering dropping the Nordic combined on the grounds that ability in cross-country skiing and ski jumping are more or less equivalent. Do you think this is reasonable? Explain your answer referring to the SAS output. (1)

   No, it is silly. It looks like the two abilities are practically independent; there is even a very small negative correlation between them. (-0.0106.) [1 mark].

   (e) Would it be better to run a PCA on the covariance matrix instead of the correlation matrix in this example? Who would be the gold medallist in that case? (2)

   No, because the two measurements are on different scales [1 mark]. The result would be dominated by cross-country and the gold medallist would just end up being whoever had the fastest Cross-country time. That's Alessandro PITTIN of Italy. [1 mark]